CLASSIFYING BREAST CANCER METASTASES ON HISTOLOGICAL IMAGES BY USING CNN

George Chen, Chao Li, Ao Hou, Weixiu Zeng, and Zeji Zhu
{george.chen, lichao, houao, zengweixiu, zhuzeji}@semptian.com
Machine Learning Lab, Semptian Co., Ltd., Shenzhen China

ABSTRACT

This paper illustrates how a quick learning team do a work of detection and classification of breast cancer metastases on histological whole slide images (WSIs) within one month. The work has three steps, 1, preparing the training data and testing data; 2, training GoogleNet [6] by using Caffe and test the model; 3, calculating the slide-level tumor size and patient-level pN stages. The result shows that GoogleNet is powerful and highly accurate annotation on training data is of vital importance.

***Index Terms*—** WSI, CNN, Caffe, GoogleNet, Breast Cancer Metastases, Histological Image

1. INTRODUCTION

Pathological diagnosing is the gold diagnose to clinical treatment to various diseases. However, the diagnose conclusion relies greatly on the personal experience and knowledge of a pathologist. How can CNN help pathologists improve the accuracy and reduce their workloads?

As a newly established team, Semptian Machine Learning Lab is excited to know that Diagnostic Image Analysis Group (DIAG) and Department of Pathology of the Radboud University Medical Center are hosting Camelyon17[1] after its successful Camelyon16[2]. The organizer has prepared rich data and useful tools for this grand challenge.

In a second, we made a decision to participate the challenge with enthusiasm. In the beginning, we set a basic goal of submitting our work before deadline and a higher goal of more confident work by using improved algorithms.

Among all participators in Camelyon16[2], we found that Dayong Wang [3] provided most details and acquired the best score. We decided to adopt GoogleNet as the main CNN method.

2. METHODS

2.1 Generation of Foreground Image

A WSI (Whole Slide Image) has a large number of pixels (97,792*221,184 pixels). However, tissues account for only 5~10% of the slide size. For efficient image processing, it is essential to distinguish tissue from WSI.

Hematoxylin - Eosin stain makes tissues on slides be colored. The colored pixels of the slide are regarded as foreground ones while white and black points are regarded as background and should be discarded.

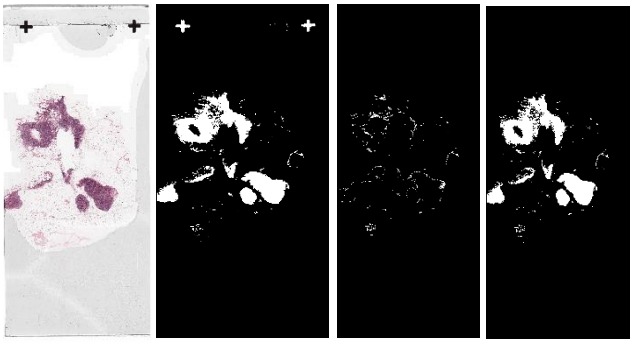
Pixels are presented in their RGB values. To get the tissue of WSIs, we convert RGB values of each pixel into HSV values [4][6][7]. We use OTSU method [5][9] to divide all pixels into foreground and background.

When using Otsu threshold to distinguish foreground we find out Otsu threshold is not best fit for some cases.

We apply Otsu method to saturation element and value element of HSV color space. For each pixel, the selection condition for foreground is

Sat>=SatOtsuThreshold and Val>=ValOtsuThreshold

We find that saturation element has two peaks for each slide, close to two ends (0.0 and 1.0). Otsu threshold for saturation element varies greatly for 0.088 to 0.46 for different slides.



(a) (b) (c) (d)

Figure 1 (a) the original image at level 5 of Tumor_001.tif; (b) the foreground image by using Otsu threshold on saturation element only; (c) the foreground image by using Otsu threshold on both of saturation and value elements; (d) the foreground image by using Otsu threshold on saturation element and artificial value (0.2) on value element;

In Figure 1, (d) shows the best result when using artificial value of (0.2) instead of Otsu threshold. (c) cuts off too much colored area.

(a) (b) (c)

Figure 2 (a) the original image at level 5 of patient_114_node_4.tif; (b) the foreground image by using Otsu threshold on saturation element and artificial value (0.2) on value element; (c) the foreground image by using artificial value (0.1) on saturation element and artificial value (0.2) on value element.

Figure 2 shows difference of using artificial value (0.1) and Otsu threshold on saturation element. In this case, the Otsu threshold is 0.44726, which is very high and leads to no foreground pixel (b).

We suggest an artificial value of 0.1 for saturation element and 0.2 for value element.

SatThreshold = min (0.1, SatOtsuThreshold)

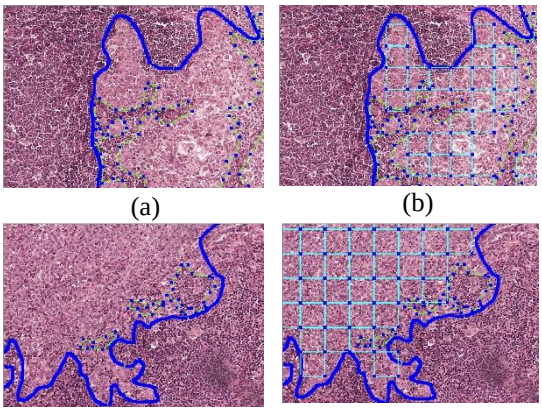
ValThreshold = min (0.2, ValOtsuThreshold)

2.2 Refined Annotation

Tumors are surrounded by negative tissue or other tissue. Most of the tumor slides in the training dataset provided by the organizer are well annotated. However, we find some problems.

- The boundary of tumor is not accurate. 256*256 pixel patches are cut off from level 0 image of slide pyramid. Some negative parts are cut off and put into positive class because of inaccurate annotation.
- Within the tumor boundary, we do find some non-positive areas which could be noise to positive class. By non-positive area we mean 1) negative parts; 2) blurred; 3) no nucleus; 4) flowing parts and 5) bubble.

The team exhaustively checks every annotated region on each slide and refined the annotations by ASAP.EXE. This way we reduce the positive patch number by about 15%.



(a) (b) (c) (d)

Figure 3 (a) part of Tumor_001.tif, blue line is the annotation provided by the organizer. Yellow polygon is the refined annotation. (b) positive patches of Tumor_001.tif. (c) another example in Tumor_014.tif. (d) positive patches of Tumor014.tif

2.3 Training CNN

We use GoogleNet in Caffe as the classification model which is a patch-based binary deep CNN model. Each patch has the size of 256*256 pixel from level 0 image. The training data comes from Camelyon16 and Camelyon17.

To each patch, we let Caffe randomly crop to 227*227 and do mirror flip

Table-1 slides for training data

	Tumor slide	**Normal slide**	**subtotal**
Camelyon16	**160**	**110**	**270**
Camelyon17	**313**	**50 (187)[1]**	**363(500)**
Subtotal	**473**	**297**	

Table -2 Training patches

	Tumor	**Non-tumor**	**Subtotal**
Patch	**188,894**	**611,361**	**800,255**

The test was done on Camelyon16 test set (130 slides). The accuracy rate on patch level is 97.6%, while TPR (true positive rate) is 80.687% and FPR (False Positive Rate) is 2.558%。

2.4 Selecting Threshold

The inference computing gives out a score between 0 and 1 for each patch per test slide. The score indicates the

[1] 50 tumor slides from 187 are annotated. We use 50 slides for training data.

probability of the patch to be positive. One slide may have several 10 thousand of patches. The conversion of patch level scores to slide level relies on the threshold. Different thresholds lead to different conclusion on the slide and tumor size.

We combine Kappa [10] value and accuracy rate to evaluate threshold when comparing to Camelyon16 test ground truth data

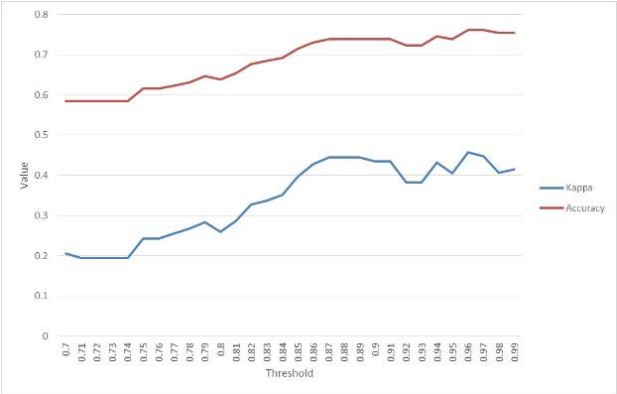


Figure 4 Kappa value changes over threshold (blue line), Accuracy rate changes over threshold (red line)

Figure 4 shows that the threshold of 0.88 has the best Kappa value (0.44) and best accuracy rate (0.76). We use this threshold to compute Camelyon17 test set.

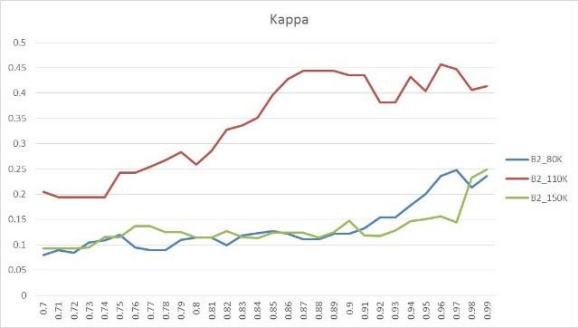


Figure 5 Kappa values for three models. (a) blue line is the model after 80K iterations; (b) red line is the model after 110K iterations; (c) green line is the model after 150K iterations

Figure 5 shows that CNN models after different number of iterations have different Kappa value. Intuitively large number of iterations leads to better accuracy. But the data shows 110K beats 150K.

2.5 Measuring Lesion

Camelyon17 requires us to measure tumors of test slides and therefore give metastases stages.

To increase the connectivity of isolated patches[11][12][13], we are allowed to expand each tumor region by 37.5 μm. One patch has size of 256*256 pixel at level 0 but it is just one pixel at level 8. To expand by 37.5 μm, we select level 7 to work so that each pixel has size of 31.1 μm, which is approximately the same as allowed expansion.

After the expansion we compute the connectivity of the region and finally use **skimage.measure.regionprops()** to get region and then use region's property **major_axis_length** to get the size of tumor.

3. CONCLUSIONS

With the completion of this project, the team concludes that

3.1 Refined annotation is very important. At the beginning, we cut patches from the annotated slide but find some patches in tumor class look like non-tumor. The team can easily recognize non-tumor portion within tumor region annotated by the organizer. We believe that non-tumor patches in the tumor class are noise and should be excluded. Training dataset should be clean. We spent several days to do the elaborative annotation and as a result the accuracy rate is much improved.

3.2 Rotation does not increase value. We increase the patch number of training dataset by rotating patch by 90°, 180° and 270° but does not see significant difference in the result.

3.3 More iterations do not give smaller lose or better accuracy.

3.4 Deep Convolutional Neural Network is an easy to use method. We are optimistic about applying CNN to pathological classification on WSI.

4. DISCUSSION

To achieve better results, the further works will focus on the following aspects.

4.1 Tissue based patch should be considered. A 256*256-pixel patch can display nucleus clearly. Should we expose the tissue or cell group to CNN? [14]

4.2 Image normalization should also be tried. A Hematoxylin-Eosin slide has simple color. LED lights and scanner sensors may disperse in RGB elements.

Stain density may also have impacts. Normalization may limit or eliminate the impact from these factors.

4.3 Looking for faster inference computing method. Each slide image contains 10k~40k patches. It takes several minutes to infer one image. it is slower than scanning time. A typical slide scanner can process one slide within 40 seconds. We would try Tensor Flow for faster inference speed.

5. REFERENCES

[1] Camelyon17.https://camelyon17.grand-challenge.org/

[2] Camelyon16. https://camelyon16.grand-challenge.org/

[3] Wang,D.,et al. "Deep Learning for Identifying Metastatic Breast Cancer". arXiv preprint arXiv:1606.05718, 2016

[4] Wikipedia, https://en.wikipedia.org/wiki/HSL_and_HSV

[5] Wikipedia, https://en.wikipedia.org/wiki/Otsu%27s_method

[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J]. arXiv preprint arXiv:1409.4842, 2014.

[7] Raphael Gonzalez, Richard E. Woods (2002) Digital Image Processing, 2nd ed. Prentice Hall Press, ISBN 0-201-18075-8, p. 295.

[8] Charles Poynton. "What are HSB and HLS?" Color FAQ. 28 November 2006。

[9] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics.9(1): 62–66,1979.

[10] Kilem Gwet (May 2002). "Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity" (PDF). Statistical Methods for Inter-Rater Reliability Assessment. 2: 1–10.

[11] Deza, Elena; Deza, Michel Marie (2009). Encyclopedia of Distances. Springer. p. 94.

[12] Samet, H.; Tamminen, M. (1988). "Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintrees". IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE. 10 (4): 579. doi:10.1109/34.3918.

[13] Michael B. Dillencourt; Hannan Samet; Markku Tamminen (1992). "A general approach to connected-component labeling for arbitrary image representations". Journal of the ACM. J. ACM. 39 (2): 253. doi:10.1145/128749.128750.

[14] Krzysztof J. Geras, et al. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. arXiv preprint arXiv:1703.07047, 2017